Computer methods to locate signals in nucleic acid sequences

Rodger Staden

MRC Laboratory of Molecular Biology, Medical Research Council Centre, Hills Road, Cambridge CB2 2QH, UK

Received 24 August 1983

ABSTRACT

This paper describes computer methods for locating signals in nucleic acid sequences. The signals include ribosome binding sites, promoter sequences and splice junctions. The methods are of use both to those trying to interpret the function of newly determined sequences and to those studying the molecular mechanisms involved in the recognition of these special signal sequences.

INTRODUCTION

We describe a computer program that can be used to locate poorly defined recognition sequences such as intron/exon junctions, ribosome binding sites and E. coli promoter sequences. Traditionally these features have been summarised as "consensus" sequences and searches have involved looking for the relevant consensus sequence. This is a rather unsatisfactory technique as it is impossible to account, except very approximately, for the relative importances of each of the bases in the sequence. Our methods for locating signals in DNA sequences assign separate values to each base at each position of the recognition sequence and can therefore indicate the relative importance of each base at each position. This is done by using a weight matrix to represent each type of recognition sequence.

A weight matrix is a two dimensional array of values that represent the score for finding each of the possible sequence characters at each position in the signal for which we are looking. For DNA sequences the weight matrix will have length equal to the length of the signal and depth of four (one row for each of A, C, G and T). Recently Stormo, Schneider, Gold and Ehrenfeucht(1) used the perceptron algorithm whose origin lies in the field of artificial intelligence to derive a weight matrix that represents the regions around prokaryotic ribosome binding sites.

We have not used perceptron algorithms but take a collection of aligned signals of the type we require and count the number of times each base occurs at each position. We then divide each of these values by the number of sequences used to compile the table to give the frequency table for the signal, and finally we calculate the natural logarithms of the frequencies. For any bases that do not occur in particular positions in any of the recognition sequences we use a value equal to the reciprocal of the number of aligned sequences that have been used to construct the frequency table: that is zero frequencies are set to this reciprocal value. The frequency tables for each of the signals shown in the figures below contain several rows of numbers: the top row (P) indicates the position in the signal, the next (N) the number of sequences that have contributed to each position in the frequency table and finally the counts for the frequency of each base (T,C,A and G) at each position. The weight matrices are the natural logarithms of the values shown in these tables.

This is our weight matrix and represents the logarithms of the probabilities of finding each base at each position in a signal. In order to locate possible signals in a sequence we simply operate on every section of the sequence with this weight matrix. This gives us a measure, for every section of the sequence, of its similarity to the collection of sequences that were used to create the weight matrix. For a signal of length L and sequence of length X this will give us X-L+1 measurements: one for each of the possible positions in the sequence at which a signal could start and be of length L.

The operation performed is the following: let the weight matrix be W(b,p), where b represents the four bases A, C, G and T, and p represents the position in the recognition sequence; and let S(j) represent the base at position j, of the sequence section being scanned which starts at position i; then we calculate the sum of W(k,p), where k is the number of the base S(j) and j=i to i+p-1. As we use logs of frequencies for the weight matrix values this is equivalent to multiplying together the relevant frequencies which went into the weight matrix. These frequencies can be thought of as probabilities of each base being part of the recognition sequence and hence their product is the probability that the section of the sequence scanned is a recognition sequence.

We apply the weight matrix to each of the sequences used in its construction and plot a histogram of the scores obtained. This histogram can show the amount of variation in the scores achieved by the sequences used to construct the weight matrix and hence give an indication of how similar they are to one another. It can also be used to decide cut-off values and scaling for the display of results when scanning new sequences.

Rather than plot a continuous line of probability for prediction of signal sequences we have chosen to display vertical lines at the positions at which the probability rises above a cut-off value and to have the height of the line represent the score. The cut-off is chosen by examining the lowest values achieved by any sequence used in the frequency table and the maximum line height is that achieved by the highest scoring sequence used to construct the frequency table. All plots are of logarithmic values.

We show examples of some of the methods below. These plots are copies taken straight from the screen of the computer terminal: all the figures, maps and scales are drawn by the program so that what is shown here is what the user of the program would see. All the sequences used in the examples have been taken from the EMBL nucleic acid sequence library(2) and the function in the program that draws the maps uses the feature tables of the library entries. To identify a feature on the map the program automatically takes the first letter of the description of the feature from the library and places it at the centre of the map line. This is why some of the map features are identified by R for reading frame, e for exon or, in the case of 16S ribosomal RNA, a letter 1 (for the beginning of 16S). Map features are drawn using alternately full and dashed lines.

RESULTS

Splice junctions

The collection of sequences used to calculate the weight matrix for splice junctions is that of Mount(3). This contains 130 different acceptor junctions aligned with the obligatory(4) AG bases and 139 donor sequences aligned with the obligatory GT bases. The frequency tables derived from these are shown in table 1.

We have decided to make the conserved AG and GT obligatory in the routines that search for these signals: only sequences that contain the obligatory bases at the appropriate positions will produce values in the plots and every position in the sequence that does contain them will, at least, be registered as a single point. The weight matrices were applied to

Table 1.

Frequency table for 3' ends of introns (acceptors)

P	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1	2	3
N	113	113	114	126	126	126	127	127	127	129	130	130	130	130	130	130	130	130
T	58	50	57	67	75	62	62	57	57	73	75	38	40	0	0	11	48	37
C	21	28	35	27	30	38	42	35	46	46	36	28	84	0	0	23	28	42
A	17	11	11	19	8	19	14	24	15	4	13	33	5	130	0	29	22	25
G	17	24	11	13	13	7	9	11	9	6	6	31	1	0	130	67	32	26

Frequency table for 5' ends of introns (donors)

P	-4	-3	-2	-1	1	2	3	4	5	6	7	8
N	139	139	139	139	139	139	139	139	139	139	136	136
T	28	10	18	17	0	139	9	16	7	87	30	36
C	42	60	16	8	0	0	3	13	3	17	28	40
A	42	56	89	12	0	0	86	94	12	23	53	33
G	27	13	16	102	139	0	41	16	117	12	25	27

all the sequences in the Mount collection and a histogram of scores calculated, as is shown in figures 1 and 2.

There are tails at the lower ends of each distribution which correspond to a few junctions that are further from the consensus than the majority. For scaling the plots we use the lowest and highest values from this

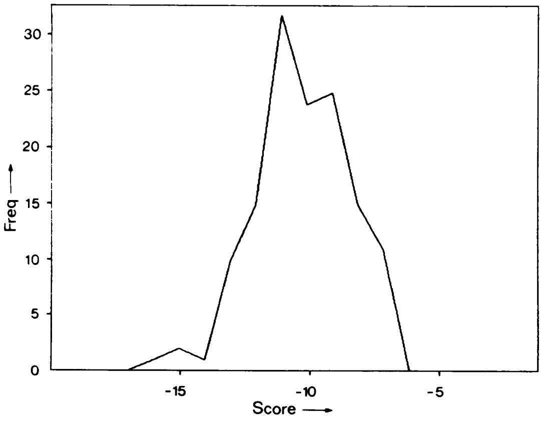


Figure 1 Histogram of scores for exon/intron junctions

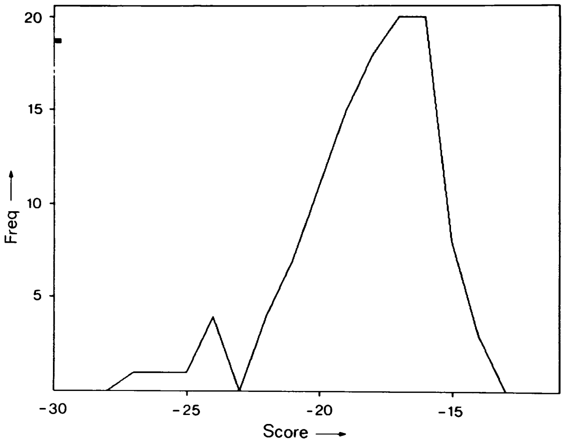
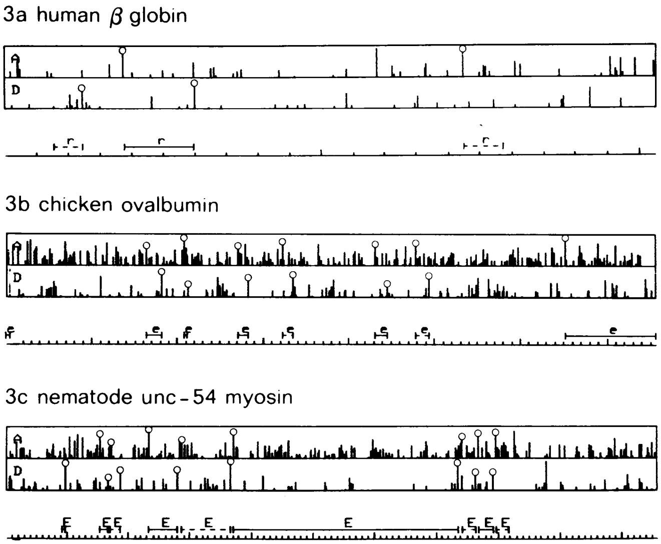


Figure 2 Histogram of scores for intron/exon junctions

histogram. Less cluttered plots, that might miss some unusual junctions, could be obtained if we treated the low scores from the tails of the histograms as special cases by raising our scaling values above them. To switch to using different cut-offs requires only that two values are changed in the program.

Examples of the application of these weight matrices to sequences where we know the positions of the intron/exon boundaries are shown in figure 3. The sequences are: in 3a. the human beta globin gene(5), in 3b. the chicken ovalbumin gene(6) and in 3c. the unc-54 myosin gene of soil nematode C. elegans(7). Each of the figures contains a plot for the acceptor predictions marked with a letter A at the left end, a plot for donor sequences marked with a letter D, a gene map showing each exon and a scale below marked in hundreds of bases. We can see that all the known junctions give a reasonably high score (these are marked with small circles at the tops of their scores), but that there many other potential splice sites some of which give even higher peaks.

The myosin sequence is about 10580 bases long and so the potential splice sites are not as closely packed as it might appear from the figure. We see that the highest scoring predictions within the local vicinity of known splice sites are those that are actually used but there are more significant peaks slightly further away within both introns and exons. For



3a human β globin

3b chicken ovalbumin

3c nematode unc-54 myosin

Figure 3 Predictions for intron/exon (A) and exon/intron (D) junctions

example in intron 1 of the myosin sequence the acceptor site is locally the highest but there are 3 other potential acceptors within the intron that score higher. In the long exon of the myosin sequence there are at least two potential donor sites that score higher than the lowest-scoring known donor (that for exon 2), but they do not score as high as the donor site that is used. Similar features are seen in the plots for the ovalbumin sequence which is 7654 bases long but the haemoglobin sequence, only 2052 bases long contains fewer "false" predictions.

Prokaryotic ribosome binding sites

We have not derived our own values for finding prokaryotic ribosome binding sites but have used the weight matrix w101, (shown in Table 2) of Stormo, Schneider, Gold and Ehrenfeucht(1). Using an algorithm that sums the appropriate values in the matrix they report that this matrix gives a score of at least 2 for all gene starts in their library whereas all other sequences score 1 or less. The weightings were derived using a perceptron algorithm. Weight matrices with shorter lengths were found to be less good despite the fact that protection experiments show that the ribosome interacts with at most 35-40 bases(8). In our application of this weight matrix we scan each reading frame separately and the results are displayed to fit

Table 2

Weight matrix W101 as used by the prokaryotic ribosome binding site search

P	-60	-59	-58	-57	-56	-55	-54	-53	-52	-51	-50	-49	-48	-47	-46	-45	-44	-43	-42	-41	-40	-39	-38	-37	-36
T	5	1	-3	9	-14	7	15	-5	3	-16	-17	4	18	5	-3	-1	2	4	5	-5	7	8	-5	-15	6
C	-21	-6	-11	-21	0	8	-7	-12	-1	1	0	-19	12	-3	-1	10	2	-8	-5	-11	8	1	23	6	-5
A	7	-2	13	-2	-8	-13	-18	5	0	-5	13	8	-15	9	-4	-7	9	0	-8	-11	-10	-6	-7	-5	-6
G	-6	-9	-7	0	8	-16	-4	-2	-16	1	-4	8	-14	5	11	-13	-24	3	7	22	-11	-9	-15	10	-4

P	-35	-34	-33	-32	-31	-30	-29	-28	-27	-26	-25	-24	-23	-22	-21	-20	-19	-18	-17	-16	-15	-14	-13	-12	-11
T	3	4	16	-4	7	11	-4	-1	12	8	10	-1	1	8	2	-10	-16	11	1	-3	16	-3	-36	-8	-27
C	2	-14	-3	-8	-10	-21	2	0	-2	-1	-11	-3	-1	5	-11	-4	7	0	-14	6	-8	-20	-7	-36	-44
A	-12	-1	-27	-3	-6	0	-12	-3	-4	-7	14	-2	-4	-6	0	12	5	-9	0	-11	-11	10	8	2	8
G	4	-5	-6	-3	-1	-4	-1	-4	-15	0	-14	3	10	-19	-3	-10	-7	-7	7	1	-8	-6	15	21	42

P	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
T	-53	-27	-26	-23	2	-7	-14	-40	-28	0	-53	75	-62	-20	-40	-10	-35	-5	-12	-1	4	14	-23	7	-2
C	-15	-50	-43	-35	-38	-29	-29	1	-9	1	-87	-55	-64	-45	11	-22	-14	-20	-15	-15	-10	-22	-5	2	6
A	0	-3	-5	4	-20	-11	5	6	-2	-15	66	-69	-52	-5	-4	6	8	-24	-7	-10	-7	13	14	-9	-18
G	35	22	16	-6	-5	-15	-25	-33	-28	-53	-36	-50	107	-5	-37	-44	-27	-15	-23	-16	-29	-47	-17	-29	-15

P	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
T	-26	1	4	-7	3	-4	0	-10	8	-18	7	-22	-21	8	4	-3	-6	7	-8	1	-5	-16	-16	7	-6
C	6	-8	19	-7	9	-3	17	-2	3	-9	5	22	22	8	-1	1	18	6	11	-10	-8	7	10	0	7
A	14	-12	-42	1	-5	-4	-32	12	-10	20	-6	-1	3	-4	4	-10	-1	-2	-14	11	14	-3	2	-13	5
G	-23	-7	-1	-6	-17	-4	0	-15	-14	-4	-17	-10	-5	-13	-8	10	-13	-13	9	-4	-3	10	2	4	-8

P	40
T	0
C	14
A	5
G	-21

conveniently with the "gene search by content" methods described elsewhere (see discussion). We give no example of this method in this paper.

Eukaryotic ribosome binding site searches

Recently Sargan, Gregory and Butterworth (9) put forward the hypothesis that there is an interaction between some mRNA leader sequences and a highly conserved structure in the 18S rRNA of eukaryotic ribosomes. The attempt to substantiate the hypothesis includes a table of base frequencies for eukaryotic sequences immediately 5' of start codons. They examined 102 sequences and we have used the base frequencies they found (see table 3) to calculate a weight matrix for searching for eukaryotic gene starts. We have not yet been able to test the method in the way we have done for the others but we include it in case the hypothesis holds and offer the program as a possible testing device. We show no example of the application of this method but have found that it does indicate the correct translation start positions for a number of sequences that were not used in the creation of the weight matrix.

E. coli promoter sequences

The frequency table we are currently using as a weight matrix to help locate promoters in E. coli sequences is taken from the compilation of such sequences recently produced by Hawley and McClure(10). E. coli promoters have been shown to contain 2 regions of conserved sequence located about 10 and 35 bases upstream of the transcription startsite(11,12,13,14). Their consensuses are TATAAT and TTGACA with an allowed spacing of 15 to 21 bases between. The spacing with maximum efficiency is 17 bases(15,16,17,18,19) and all but 12 of the 112 sequences in the Hawley and McClure collection could be aligned with a separation of 17 +or-1 bases. The spacing between the -10 region and the startsite is usually 6 or 7 bases but varies between 4 and 8 bases. There is an AT rich region of 8 to 10 bases upstream of the -35 region. Hawley and McClure also show a conserved section to exist around the +1 region. The frequencies for the three regions are shown in table 4.

For our search we have taken the logarithms of the frequency tables for the three conserved sections of sequence (the -35, -10 and +1 regions) as

Table 3

Frequencies for eukaryotic ribosome binding sites

P	-7	-6	-5	-4	-3	-2	-1	1	2	3
	102	102	102	102	102	102	102	102	102	102
T	19	24	31	12	0	18	5	0	102	0
C	20	15	32	65	5	42	52	0	0	0
A	50	27	27	19	86	36	34	102	0	0
G	6	29	12	6	11	6	11	0	0	102

Table 4.

Frequency tables for E. coli promoters

-35 region:

P	-50	-49	-48	-47	-46	-45	-44	-43	-42	-41	-40	-39	-38
N	107	109	109	110	110	110	110	110	110	111	111	110	111
T	41	33	32	25	34	22	35	35	42	27	32	42	47
C	22	27	18	29	20	14	20	12	22	23	16	25	10
A	28	38	30	37	35	56	42	42	37	42	39	18	25
G	16	11	29	19	21	18	13	21	9	19	24	26	29

-35 region continued:

P	-37	-36	-35	-34	-33	-32	-31	-30	-29	-28	-27	-26
N	112	112	112	112	112	112	112	112	112	112	112	112
T	14	92	94	11	19	15	37	46	34	38	48	34
C	43	7	6	11	18	60	8	25	23	23	17	20
A	26	2	6	2	72	26	50	26	34	25	26	31
G	29	11	6	88	3	11	17	15	21	26	21	27

-10 region:

P	-23	-22	-21	-20	-19	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5
N	112	112	112	112	112	112	112	112	112	112	112	112	112	112	112	112	112	112	112
T	35	28	28	27	39	51	34	43	26	31	89	3	49	15	19	108	31	29	21
C	34	21	24	27	12	25	20	25	20	27	10	2	16	14	22	3	13	16	30
A	20	39	33	33	39	23	29	16	23	19	2	106	29	66	57	1	35	23	31
G	23	24	27	25	22	13	29	28	43	35	11	1	18	17	14	0	33	24	30

+1 region:

P	-2	-1	1	2	3	4	5	6	7	8	9	10
N	86	88	88	88	88	88	88	88	88	88	88	88
T	16	22	2	42	27	23	20	25	27	15	16	29
C	29	49	4	25	25	13	18	22	17	17	16	17
A	20	9	45	16	24	25	28	24	24	32	35	26
G	21	8	37	5	12	27	22	17	20	24	21	16

calculated from the Hawley and McClure alignment and look for each in turn. We look first for a sufficiently high scoring -35 sequence, then for the best -10 sequence above a minimum score within the allowed distances from the -35, then for the best +1 sequence above a minimum score within range of the -10 sequence. If all three can be found a vertical line is drawn, the height of which represents the overall score for the three sections of sequence. To examine the similarity of all the sequences in the Hawley and McClure collection to the average sequence, and to determine the cutoff values for plotting of results we applied the weight matrix to each of them. The results shown as a histogram of scores can be seen in figure 4. These are logarithmic values and so we can see that there are large variations in the sequences: the range being -60.4 for recA(20,21) to -77.3 for rpoB(22,23).

The sequences that lie to the left of the local minimum at about -72. are those for minor promoters such as rpoB, those of low activity such as lacI(24,25) or those that require accessory proteins for their activity such as lambda PRE(26,27) The cut-off values for each of the three regions are the lowest values found for all of the sequences in the Hawley and McClure collection. As is mentioned above the efficiency of promoters with varying gaps between the -35 and -10 regions has been measured and to take this into account we apply gap penalties to the plots. The values used are shown below.

Gap penalties for -35 to -10:

15	0.02	(only exists as mutant(18))
16	0.2	
17	1.0	
18	0.2	
19	0.05	(guess)
20	0.02	(guess)
21	0.01	(guess)

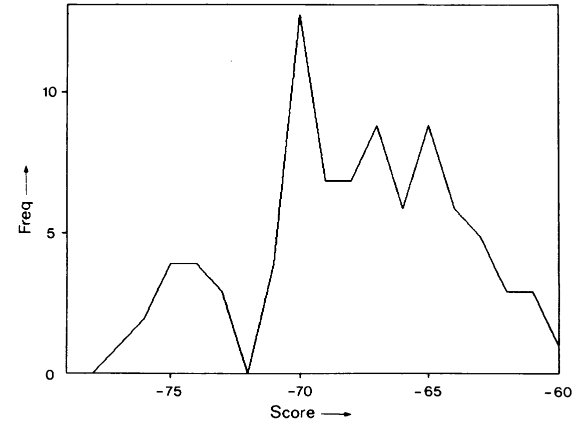


Figure 4 Histogram of scores for E.coli promoters

In practice we draw 5 plots for any section of sequence: 1) all -35 regions that score above a certain cut-off; 2) all -10 regions that achieve a score above a certain cut-off; 3) all promoter-like sequences that contain all three conserved regions above the relevent cut-offs and with the allowed gap distances, but applying no gap penalties; 4) as for 3 but applying gap penalties for the distance between the -35 and -10 regions; 5) all promoter-like sequences on the complementary strand of the sequence. The cut-offs and scaling for the seperate -35 and -10 searches were determined by calculating the range of values observed for known promoters and then extending the range by + and - 10% so that for example the lowest -35 plotted will score 10% below the lowest known -35 region and the top of scale is 10% higher than any known -35 region reaches. The other plots have cut-offs and are scaled using the actual values observed for the known promoters in the Hawley and McClure collection. Examples of the plots except those for the complementary strand are shown in figure 5.

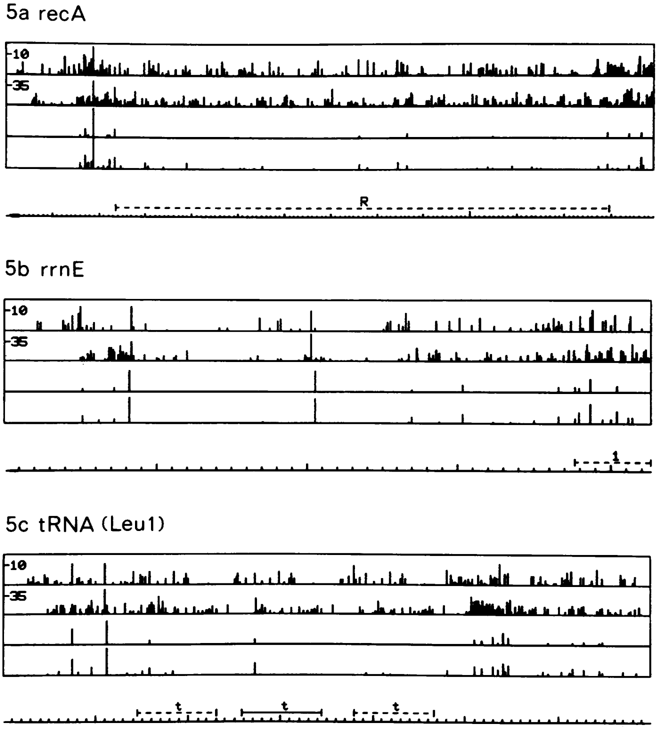


Figure 5 Predictions for promoters in E. coli

These plots are for recA(20), rrnE(28) and leu1 tRNA(29). Each plot includes a gene map with a scale marking every tenth base below. For the recA sequence we see that there is a very clear peak in the correct position(21), that the gap size is not optimal (the plot with gap penalties is not the same height as that without), and that both the -35 and -10 regions score highly. We also see that although there are many separate (-35) and (-10)-like sequences elsewhere there are very few that are in the correct relative positions to one another and therefore there are no striking predictions for promoters in the rest of the sequence. This protein is produced at a low basal level during normal cell growth but is formed more efficiently after treatments that cause damage to DNA. These facts and the high score for this promoter are in agreement with the idea that this gene is inhibited by another protein (thought to be lexA(see references in 21)).

The results for rrnE show two strong peaks corresponding to the postions of the two known transcription startsites 283 and 174 basepairs upstream from the 16S mrRNA coding sequences(30). There are no other peaks of comparable size but a number of separate (-35) and (-10)-like sequences.

Examination of the plot for the tandem leu1 tRNA sequences shown in figure 5c. shows that there is a high peak with a subsidiary peak 40 basepairs upstream. The plots for the separate -35 and -10 regions in the area corresponding to this upstream peak indicate that the major contributor to this high score is the -10 sequence. There is a single peak in the second tRNA gene and a cluster of weak peaks downstream of the third tRNA.

DISCUSSION

The routines described will, with varying degrees of success, indicate the most probable locations of various signals in nucleic acid sequences. The fact that they do not unambiguously identify the signals has a number of possible explanations. Clearly the method used has little correspondence to the physical processes that must be involved when enzymes recognize their respective binding sites. No account is taken of possible cooperativity, both positive and negative, between the various parts of a signal sequence: any cooperativity would be lost by the averaging effect of compiling the frequency table. No account is taken, except for the fairly clear cut cases of promoter sequences, of the possibility of insertions and deletions in the signal sequences: the frequency tables are constructed by straight alignment starting from obligatory or highly conserved bases. No account is taken of the possibilty that these signal sequences fall into separate classes and that if so separated they would look far more alike than they do when all compiled together. This last point requires further investigation and may result both in a greater understanding and better predictive programs.

The scaling of the plots has been chosen by applying the weight matrices to the sequences used in their construction. It should be noted therefore that any section of sequence that produces a definite value on any of the plots is as similar to the "consensus" as at least one of the sequences that were used to construct the weight matrices and hence could be viewed either as a possible signal or as a false prediction.

Despite all these points against the use of weight matrices for locating signal sequences the methods are so far, the best we can do, and as we have seen in the examples work quite well. The methods also serve the purpose of indicating to those using the program the often large number of possible alternatives to their favourite location for a signal sequence: often when using "eyeballing" techniques the first reasonable consensus found in the right region will be taken as being the only possible alternative.

As soon as sufficient data are available we will add further searches, such as for eukaryotic promoter sequences. Now that the overall framework for the use of weight matrices is set up in the program it is very easy to make additions or changes. We might also consider using different cut-off and scaling values based on better measures of variation than the range.

We originally used weight matrices in 1979 to investigate possible sliding mechanisms for RNA splicing (R. Staden and G. G. Brownlee, unpublished). Better experimental techniques are now available to investigate the predictions and so the methods described may thus now be more usefully employed for this purpose. The different searches will indicate signals not known to be used that are of equal "strength" to those that are. In this way the program can be used to draw the experimenter's attention towards "false" predictions of signals: are they infact used, or are they different; are they masked in some way; do they act indirectly in a positive way by increasing local concentrations of enzymes, or in a negative way by competing for available enzymes? For the case of the eukaryotic ribosome binding site search, the usefulness of which we have not systematically tested, the program could be used to examine the hypothesis presented in reference(9), of an interaction between 18S ribosomal RNA and the 5' leader sequence of eukaryotic mRNA's.

We have previously described methods that locate genes by looking at the effects that coding for a protein has on a DNA sequence(31). We refer to these techniques that examine the content of coding regions as "gene search by content" methods and they are, so far, more likely to correctly indicate coding regions than any of the "gene search by signal" methods described in the current paper. Using various collections of aligned sequences we have been employing weight matrices to find signals for several years but until we developed the "gene search by content" methods we considered that, if used for locating genes, the techniques were not worth describing in a formal publication. However, the combination of the two types of search and the other functions included in the program that contains these methods now gives a very powerful way of analysing sequences to find genes.

This program called ANALYSEQ, which will be described elsewhere, works on a simple graphics terminal(32) and has the ability to superimpose the results of all the different types of search. When the different types of search, which will often also be independent of one another, are superimposed we can get a much clearer interpretation of the function of newly determined DNA sequences.

While this manuscript was in preparation a paper by Harr, Haggstrom and Gustafsson(33) which describes methods for locating signals employing weight matrices was published. Instead of using frequencies to construct the weight matrices they use values that are divided by the highest count in each column of the table.

ACKNOWLEDGEMENTS

I am grateful to A. A. Travers for information and advice about promoter sequences and to A. D. McLachlan and F. Sanger for critical reading of this manuscript.

REFERENCES

(1) Stormo, G. D., Schneider, T. D., Gold, L. and Ehrenfeucht A., (1982) Nucl. Acid Res. 10, 2997-3011.
(2) EMBL Nucleotide Sequence Data Library, European Molecular Biology Laboratory, Postfach 10 22 09, D-6900 Heidelberg, West Germany.
(3) Mount, S. M. (1982), Nucl. Acid Res. 10 459-472.
(4) Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P, (1980) Cell 20, 625-637.
(5) Lawn, R. M., Efstratiadis, A, O'Connell, C and Maniatis, T, (1980) Cell 21 647-651.

(6) Woo S. L. C., Beattie W. G., Catterall J. F., Dugaiczyk A., Staden R., Brownlee G. G. and O'Malley B. W. (1981), Biochem. 20, 6437-6446.

(7) McLeod, A. R., Karn, J. and Brenner, S. (1981) Nature 291, 386-390.

(8) Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B. S. and Stormo, G (1981), Ann. Rev. Microbiol. 35, 365-403.

(9) Sargan, D. R., Gregory, S. P. and Butterworth, P. H. W. (1982) Febs Let. 147 133-136.

(10) Hawley, D., K. and McClure, R., (1983) Nucl. Acid Res 11, 2237-2255

(11) Pribnow, D. (1975) J. Mol. Biol. 99, 419-443.

(12) Takanami, M., Sugimoto, k., Sugisaki, H. and Okamato, T. (1976) Nature 260, 297-302.

(13) Schaller, H., Gray, C. and Herrman, K. (1975) Proc. Natl. Acad. Sci. 72, 737-741.

(14) Seeburg, P. H., Nusslein, C. and Schaller, H. (1977) Eur. J. Biochem. 74, 107-113.

(15) Jaurin, B., Grundstrom, T., Edlund, T. and Normark, S. (1981) Nature 290, 221-225.

(16) Stephano, J. E. and Gralla, J. D. (1982) Proc. Natl. Acad. Sci. 79, 1069-1072.

(17) Youderian, P., Bouvier, S. and Susskind, M. (1982) Cell 30, 843-853.

(18) Berman, M. L. and Landy, A. (1979) Proc. Natl. Acad. Sci. 76, 4303-4307.

(19) Mandecki, W. and Reznikoff, W. S. (1982) Nucl. Acid Res. 10, 903-912.

(20) Sancar, A., Stachelek, C., Konigsberg, W. and Rupp, W., D. (1980) Proc. Natl. Acad. Sci. U.S.A. 77, 2611-2615.

(21) Horii, T., Ogawa, T. and Ogawa, H. (1980), Proc. Natl. Acad. Sci. 77, 313-317.

(22) Post, L. E., Strycharz, G. D., Nomura, M., Lewis, H. and Dennis, P. P. (1979) Proc. Natl. Acad. Sci 76, 1697-1701.

(23) An, G. and Friesen, J. D. (1980) J. Bacteriol. 144, 904-916.

(24) Calos, M. (1978) Nature 274, 762-765.

(25) Steege, D. A. (1977) Proc. Natl. Acad. Sci. 74, 4163-4167.

(26) Rosenberg, M., Court, D., Shimatake, H., Brady, C. and Wulff, D. L. (1978) Nature 272, 414-423.

(27) Schmeissner, U., Court, D., Shimatake, H. and Rosenberg, M. (1980) Proc. Natl. Acad. Sci. 77, 3191-3195.

(28) de Boer, H., A., Gilbert, S., F. and Nomura, M. (1979) Cell 17, 201-209.

(29) Duester, G., Camper, R., K. and Holmes W., M. (1981) Nucl. Acid Res. 9, 2121-2139.

(30) Gilbert, S., F., de Boer, H., A. and Nomura, M. (1979), Cell 17, 211-224.

(31) Staden, R. submitted

(32) The programs are written in FORTRAN 77 for a VAX 11/780 computer manufactured by Digital Equipment Corp., Maynard, Mass. USA. The graphics terminal we use is a Retro-graphics VT640 which is an enhancement to the VT100 computer terminal made by DEC. The enhancement is made by Digital Engineering Inc., Sacramento, Calif. USA.

(33) Harr, R., Haggstrom, M. and Gustafsson, P. (1983) Nucl. Acid Res. 11, 2943-2957.